How will historians explain AI? – Mark Bell, Digital Researcher, The National Archives

I see there being two types of explainability of a machine learning system: a description of the algorithm itself, how it functions, and how well it performs; or an explanation of individual decisions that the system has made. It seems to me that most of the effort in producing explanations interpretable by the average person goes into the second challenge, and less so the first. I'm considering the field from the perspective of the archive and how an archived machine learning model may be studied in, say, 100 years from now. Rather than considering a model that has been preserved for historical value I will focus on preservation for the purpose of holding some process to account at a later date. What should we therefore record now in order to understand, and explain decisions made by this algorithm in the future?

Traditional computing systems are built around deterministic rules, elicited during requirements workshops and codified by architects and designers. In order to understand a system's behaviour the documentation is vital. We can analyse the outputs to assess how well those rules have been implemented by creating test data which follows all of the branches of logic defined within the documentation. Machine Learning turns this process on its head. First data is created, or more likely acquired, and then a system is built to model that data as correctly as possible. Modern deep learning approaches require massive volumes of data but that isn't always a problem, it is understanding vast quantities of complex data in all its nuance which is. Two books, Weapons of Maths Destruction[i], and Algorithms of Oppression[ii], catalogue the consequences of using high volumes of poorly understood or inappropriate, proxy, data to drive decisions. So it seems preserving the training data is of paramount importance, but this quickly becomes more complicated than it sounds. Consider a system which learns over time, perhaps using a Bayesian approach to update its parameters as it is used. What was the state of the data at the point any particular decision was made? How can we summarise the data over the lifetime of the system? Or what about a system which uses transfer learning, where a pre-trained network is used as a starting point for a model trained on domain specific data. The original data that was fed into the pre-trained network may well be unavailable.

After the data comes the design process. What exploratory analysis was performed? How did the results influence the subsequent model design? Government departments are now encouraged to use the Reproducible Analytical Pipeline [iii](RAP), designed by the Government Digital Service (GDS), which provides a framework for producing official statistics which are fully reproducible. This is a positive step and this approach can be used as a way of capturing the thought process for the design of a machine learning model. Perhaps design is the wrong word. Google AI researcher Ali Rahimi, at NIPS 2017[iv], eloquently described the design of deep learning models as being similar to alchemy. He argued that the building blocks of these models are often used because they seem to produce desired results rather than because of a deep theoretical understanding of their function. He describes machine learning research as a competition, aiming to beat previous benchmarks on standardised datasets, rather than a research activity. In a subsequent paper[v], he and his collaborators suggested a set of standards for empirical evaluation of new models. Some of the suggestions address the issue that very complex models can produce excellent results but in fact it may be one component which is responsible for most of the performance and in fact a simpler model would have sufficed. Frankle and Carbin [vi]describe the training of large neural networks as like buying many lottery tickets. They propose a method for identifying smaller sub-networks, the so called winning tickets, which not only provides efficiency in training but sounds like a step towards interpretability.

More classical learning approaches set a standard to aim for, but also suggest it may be a long run to full interpretability. The techniques used in Linear Regression have been around for over 200 years, and a whole set of robust and theoretically sound tools have been developed to understand their performance and accuracy. We should expect that decades from now we will have similar levels of understanding of modern algorithms. This, perhaps optimistic viewpoint given the complexity of modern algorithms, is expressed by Hastie and Efron [vii]who provide a timeline demonstrating how theoretical understanding has always lagged behind applied methods; a sentiment echoed by Yann LeCun in his response to Al Rahimi's speech. Hastie and Efron argue that statistics in the 19th century revolved around applied techniques which seemed to work and it took until the mid-20th century to get it onto a firm mathematical footing. Since then the field has moved in a computational direction and once again further away from mathematical theory but there are signs of the statistical community catching up again.

Finally, if we really want to put an algorithm in context and understand its function then we need to capture the decisions it made. Its actions in the real world are where we truly hold it to account. Performance in against benchmarks, confidence bands, F-scores, and the like, become meaningless once the model is out in the wild making life affecting decisions. Only then can we assess its fairness and accuracy. However, at this point we come back full circle to the training of the algorithm. Similarly to holding a human decision making process to account, where we may look at whether the people involved had enough information and knowledge to make decisions, can we do the same with computer algorithms? Is it possible to create a five star scale[viii], as Berners-Lee devised for Open Data, for long term explainability accounting for training data, algorithm design/selection, and evaluation?

[i] Cathy O'Neil. 2016. Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. Crown Publishing Group, New York, NY, USA.

[ii] *Noble, S. (2018). Algorithms of Oppression: How Search Engines Reinforce Racism. New York: NYU Press*

[iii] https://dataingovernment.blog.gov.uk/2017/03/27/reproducible-analytical-pipeline/

[iv] *Viewed at https://www.youtube.com/watch?v=Qi1Yry33TQE*

[v] Sculley, D., Snoek, J., Wiltschko, A., & Rahimi, A. (2018). Winner's curse? On pace, progress, and empirical rigor.

[vi] *Frankle, J., & Carbin, M. (2018). The lottery ticket hypothesis: Finding sparse, trainable neural networks. arXiv preprint arXiv:1803.03635.*

[vii] Bradley Efron and Trevor Hastie. 2016. Computer Age Statistical Inference: Algorithms, Evidence, and Data Science (1st ed.). Cambridge University Press, New York, NY, USA.

[viii] *https://www.w3.org/2011/gld/wiki/5_Star_Linked_Data*